

Statistical Methods for Artificial Intelligence, Autumn 2006
Problem set 6, Due Wednesday Nov. 22

Problem 1. This problem is on Boosting. Consider the problem of finding a sequence of decision trees f_1, \dots, f_N where each tree takes $x \in \mathcal{X}$ and returns either 1 or -1 and also want a series of coefficients β_1, \dots, β_N . Let F be the sequence of trees and let β be the sequence of coefficients. We have the following notation.

$$\sum_{i=1}^N \beta_i f_i(x) = \beta \cdot F(x)$$

Ideally we would like to find both F and β as follows where this can be viewed as an instance of the general regression equation with log loss and with $\lambda = 0$ but where we are simultaneously searching over possible features (the “weak learners”).

$$\langle F^*, \beta^* \rangle = \operatorname{argmin}_{F, \beta} \sum_{t=1}^T \ln(1 + \exp(-y_t \beta \cdot F(x_t)))$$

In Boosting we are interested in optimizing a loss function by a kind of “gradient descent”. Assume that we have constructed β and F but now want to extend it one more tree f_{N+1} and one more weight β_{N+1} . We will construct the tree f_{N+1} by calling a decision tree learner on a reweighting of the training data. Give an expression for the weight of the training point $\langle x_t, y_t \rangle$ used in the gradient descent version of Boosting. Your expression should be given in terms of β and F and the training data.

Problem 2. This problem is on convexity for log loss in the structured case. In the structured case with logistic loss we want to solve the following

optimization problem.

$$\begin{aligned}\beta^* &= \operatorname{argmin}_{\beta} \left(\sum_{t=1}^T \ln(1/P(y_t|x_t, \beta)) \right) + \lambda \|\beta\|^2 \\ &= \operatorname{argmin}_{\beta} \left(\sum_{t=1}^T \beta \cdot \Psi(x_t, y_t) + \ln Z(x_t, \beta) \right) + \lambda \|\beta\|^2\end{aligned}$$

$$Z(x_t, \beta) = \sum_{y \in \mathcal{Y}} \exp(\beta \cdot \Psi(x_t, y))$$

The right hand side is a function of β . This function of β is smooth (it is continuous and infinitely differentiable). A smooth function of vector space is convex if and only if for every point in the space, and every vector direction from that point, the second derivative of the function in that direction is non-negative. This can be formulated mathematically in terms of the Hessian matrix. The Hessian H of a function $f(\beta)$ is defined as follows.

$$H_{i,j} = \frac{\partial^2 f}{\partial \beta_i \partial \beta_j}$$

In general $\alpha^T H \alpha$ is the rate of change in the direction α of the derivative of f in the direction α . If the Hessian exists then f is convex if and only if the Hessian is positive semidefinite, i.e., for any direction α the second derivative in the direction α , which equals $\alpha^T H \alpha$, is non-negative.

a. Compute the Hessian matrix of the function $\|\beta\|^2$. Show that $\|\beta\|^2$ is convex.

b. Show that the Hessian of $\ln Z(x_t, \beta)$ is the covariance matrix of the vectors $\Psi(x_t, y)$ weighted by $P(y|x_t, \beta)$. Explain why this implies that $\ln Z(x_t, \beta)$ is convex in β .

c. Explain why parts a. and b. imply that the above optimization problem is convex in β .

Problem 3. (Understanding L1 and L2 norms) The L2 norm of a vector $x \in \mathcal{R}^n$ is defined as $\|x\|_2 = \sqrt{\sum_i x_i^2}$ and the L1 norm is $\|x\|_1 = \sum_i |x_i|$.

a. For $x, y \in \mathcal{R}^n$, use the Cauchy-Schwarz inequality, which states that $x \cdot y \leq \|x\|_2 \|y\|_2$, to show that:

$$\|x\|_2 \leq \|x\|_1 \leq \sqrt{n} \|x\|_2$$

b. If x lives in the unit sphere (i.e. $\|x\|_2 = 1$), show that both the upper and lower bounds are achievable with points in the sphere. Namely, show there exists an x in the sphere such that $\|x\|_1 = 1$ and that there exists an x in the sphere where $\|x\|_1 = \sqrt{n}$. Hence, the L1 diameter of the sphere can be quite large, \sqrt{n} .

c. If x lives in the simplex (meaning that all the coordinates are positive and sum to one, i.e. it is a probability distribution), again show that the upper and lower bounds are achievable with points in the simplex. Hence, this implies the L2 diameter of the simplex is bounded by 1.